Active Learning for Black-Box Adversarial Attacks in EEG-Based Brain-Computer Interfaces

Xue Jiang, Xiao Zhang and Dongrui Wu
School of Artificial Intelligence and Automation
Huazhong University of Science and Technology, Wuhan, China
Email: xuejiang@hust.edu.cn, xiao_zhang@hust.edu.cn, drwu@hust.edu.cn

Abstract—Deep learning has made significant breakthroughs in many fields, including electroencephalogram (EEG) based brain-computer interfaces (BCIs). However, deep learning models are vulnerable to adversarial attacks, in which deliberately designed small perturbations are added to the benign input samples to fool the deep learning model and degrade its performance. This paper considers transferability-based black-box attacks, where the attacker trains a substitute model to approximate the target model, and then generates adversarial examples from the substitute model to attack the target model. Learning a good substitute model is critical to the success of these attacks, but it requires a large number of queries to the target model. We propose a novel framework which uses query synthesis based active learning to improve the query efficiency in training the substitute model. Experiments on three convolutional neural network (CNN) classifiers and three EEG datasets demonstrated that our method can improve the attack success rate with the same number of queries, or, in other words, our method requires fewer queries to achieve a desired attack performance. To our knowledge, this is the first work that integrates active learning and adversarial attacks for EEG-based BCIs.

Keywords—brain-computer interfaces; adversarial examples; active learning; black-box attack

I. INTRODUCTION

A brain-computer interface (BCI) is a communication system that connects a human brain and a computer [1]. Direct dialogue between the brain and the computer can be achieved by the established information path. Electroencephalogram (EEG) is the most frequently used input signal in BCIs, due to its low cost and convenience [2]. Various paradigms are used in EEG-based BCIs, such as P300 evoked potentials [3]–[6], motor imagery (MI) [7], steady-state visual evoked potential (SSVEP) [8], etc.

Deep learning has achieved great success in numerous fields. Multiple convolutional neural network (CNN) classifiers have also been proposed for EEG-based BCIs. Lawhern *et al.* [9] proposed EEGNet, which can be applied to different BCI paradigms. Schirrmeister *et al.* [10] designed a deep CNN model (DeepCNN) and a shallow CNN model (ShallowCNN). In addition, there were some studies to convert EEG signals into images and then classify them with deep learning models [11]–[13]. This paper considers only CNN models (i.e., EEGNet, DeepCNN, ShallowCNN) which take the raw EEG signals as the input.

Despite their state-of-the-art performance, recent studies

have shown that deep learning models are vulnerable to adversarial examples, which are crafted by adding small imperceptible perturbations to benign examples to degrade the performance of a well-trained deep learning model. For example, in face recognition, an adversarial perturbation can be attached to the glasses, and the attacker who wears it can avoid being recognized, or be recognized as another person [14]. In image classification, the adversarial examples can fool a deep learning model to give incorrect image labels [15]–[18]. Many adversarial attacks have also been performed in speech recognition [19], malware classification [20], semantic segmentation [21], etc. Recently, Zhang and Wu [22] verified that adversarial examples exist in EEG-based BCIs and proposed several adversarial attack approaches.

Many effective algorithms for generating adversarial examples have been proposed, such as the fast gradient sign method (FGSM) [15], the C&W method [23], L-BFGS [16], the basic iterative method [17], DeepFool [24], etc. These methods mainly considered the white-box attack scenario, where the attacker has full access to the target model, including its architecture and parameters. Accordingly, the attacker can perform attacks by adding perturbations along the direction calculated by gradient-based strategies or optimization-based strategies. However, the white-box setting requires full information of the target model, making it impractical in many real-world applications.

In this paper, we focus on a more realistic and challenging black-box attack scenario, where the attacker can only observe the target model's responses to inputs but has no information about its architecture, parameters, and training data. The attacker needs to generate adversarial examples whose perturbations are constrained to a magnitude threshold with limited queries. Papernot *et al.* [25] proposed a black-box attack method which generated adversarial examples for a white-box substitute model and attacked the black-box target model based on the transferability. Zhang and Wu [22] proposed an unsupervised fast gradient sign method (UFGSM) to craft adversarial examples for black-box attacks in EEG-based BCIs. However, these transferability-based approaches suffered from low query efficiency: they usually require a large number of queries to build a substitute model that is sufficiently similar to the target model.

To address this issue, we introduce a query synthesis based active learning strategy to transferability-based black-

box attacks of EEG-based BCIs. Given a small amount of initial training EEG epochs for the substitute model, we first randomly obtain a pair of opposite instances in each iteration, i.e., two instances of different classes. According to the initial opposite-pair, we use a binary search strategy to synthesize another opposite-pair close to the current classification boundary in the input space. After that, we synthesize queries along the perpendicular line of the previously found opposite-pair. This query synthesis based active learning strategy can directly synthesize queries which are close to the target model's decision boundary and well scattered. It improves the query efficiency by directly searching for informative examples in the input space for substitute model training, instead of taking a fixed step along the gradient direction. Experiments on three CNN classifiers and three BCI datasets demonstrated the effectiveness of our proposed method.

The remainder of this paper is organized as follows: Section II introduces related work on black-box adversarial attacks and active learning. Section III proposes our query synthesis based active learning approach for crafting adversarial examples for black-box attacks in EEG-based BCIs. Section IV evaluates the attack performance of our proposed approach. Finally, Section V draws conclusion.

II. RELATED WORK

In this section, we briefly review previous studies on black-box attacks and active learning.

A. Black-Box Attacks

Black-box attacks can be roughly divided into three categories: decision-based, score-based and transferability-based. Decision-based attacks were first proposed by Brendel *et al.* [26]. Its main idea is to gradually reduce the magnitude of the adversarial perturbation while ensuring its effectiveness. Score-based attacks rely on the model's output scores, e.g., class probabilities or logits, to estimate the gradients and then generate adversarial examples [27], [28]. Transferability-based attacks were first proposed by Papernot *et al.* [25] in image classification. The attacker trains a substitute model, which solves the same classification problem as the target model, to generate adversarial examples for the target model.

In transferability-based black-box attacks, the key step is to learn a substitute model whose decision boundary resembles the target model's. Papernot *et al.* [25] used Jacobian-based dataset augmentation to synthesize a substitute training set and labeled it by querying the target model. By alternatively augmenting the training set and updating the substitute model, it can gradually approximate the target model. Recently, Zhang and Wu [22] extended this idea to EEG-based BCIs, but their approach was slightly different: they synthesized a new training set by using the loss computed from the inputs, instead of the labels from the target model, to calculate the Jacobian matrix.

Albeit the outstanding attack performance of these methods, they always require a large number of queries to train a substitute model. Generally, the number of queries grows exponentially with the number of iterations. In order to improve the query efficiency, we propose an active learning based data augmentation approach to train the substitute model in transferability-based black-box attacks.

B. Active Learning

Active learning is an effective way to reduce the data labeling effort, by actively selecting the most useful instances to label. There are two main scenarios of active learning in the literature: query synthesis [29]–[32] and sampling. The latter can be further divided into stream-based sampling [33], [34] and pool-based sampling [35]–[38]. Sampling-based active learning selects real unlabeled instances from a pool or steam for labeling. In query synthesis based active learning, one can query any data instance in the input space, including synthesized instances.

Intuitively, query synthesis can be applied to the training process of the substitute model to improve the query efficiency of black-box attacks, because it can actively synthesize more informative EEG epochs than generating some epochs in Jacobian-based way. Furthermore, compared with sampling, it is more efficient to synthesize a query directly instead of evaluating every instance in an unlabeled data pool. Here we don't need to worry about the fact that some synthesized EEG epochs may be unrecognizable to human [39], because the target model can label any instance in the input space. Our ultimate goal is to identify the decision boundary of the target model by a minimum number of queries.

III. METHODOLOGY

In this section, we first introduce the transferability-based black-box attack setting for CNN classifiers in EEG-based BCIs, and then describe the query-synthesis-based active learning strategy for training the substitute model, and adversarial example crafting for the target model.

A. Attack Setting

The attack framework in this paper is the same as our previous work [22], where the attackers can add adversarial perturbations before the machine learning modules.

Let $\mathbf{x}_i \in \mathbb{R}^{C \times T}$ be the i-th raw EEG epoch ($i = 1, ..., n$), where C is the number of EEG channels and T the number of the time domain samples. Let $f_\theta(\mathbf{x}_i) \to y_i$ denote the CNN model that predicts the label for an input EEG epoch.

Given a target CNN model and a normal EEG epoch $\mathbf{x}_i$, the task is to generate an adversarial example $\mathbf{x}_i^*$ misclassified by the CNN model. Formally, an adversarial example $\mathbf{x}_i^*$ should satisfy the following constraints:

$$f_\theta(\mathbf{x}_i^*) \neq y_i, \tag{1}$$

$$D(\mathbf{x}_i, \mathbf{x}_i^*) \leqslant \epsilon, \tag{2}$$

where $D(\cdot, \cdot)$ is a distance metric, and ϵ constrains the magnitude of the adversarial perturbation. (1) ensures the success of the attack, and (2) ensures that the perturbation is not larger than a predefined upper bound ϵ.

Next we describe an algorithm to learn a substitute model for a given target CNN classifier by querying it for labels on query-synthesized inputs, and then introduce the UFGSM approach [22] to craft adversarial examples on the trained substitute model, which can also be transferred to the target model. We start with binary classification, and then extend it to multi-class tasks.

B. Query-synthesis-based Dataset Augmentation for Training the Substitute Model

In the black-box attack scenario, we have no access to the architecture, parameters and training data of the target model, but can input EEG trials to the target model and observe the corresponding outputs to probe the model. As in [25], we also adopt the transferability-based approach to implement the black-box attack. The difference is that we use a query-synthesis-based active learning strategy as the data augmentation technique in substitute model training rather than the Jacobian-based approach. The oracle in query synthesis active learning is the target model.

1) Binary search synthesis: Assume a small labeled training set S_0 has been obtained from querying the target model. An initial substitute model f'_0 can be trained on this set. Suppose $\{\mathbf{x}_0^+, \mathbf{x}_0^-\}$ is an opposite-pair in S_0. Then, we query their middle point on the substitute model to find another opposite-pair closer to the decision boundary, using Algorithm 1.

Algorithm 1: Binary search synthesis. $\{\mathbf{x}^+, \mathbf{x}^-\}$ = $BinarySearch(\{\mathbf{x}_0^+, \mathbf{x}_0^-\}, f', m)$

Input: $\{\mathbf{x}_0^+, \mathbf{x}_0^-\}$, initial opposite-pair of EEG epochs; f', current substitute model; m, maximum number of binary search iterations.
Output: $\{\mathbf{x}^+, \mathbf{x}^-\}$, an opposite-pair of EEG epochs.

$\mathbf{x}^+ = \mathbf{x}_0^+$;
$\mathbf{x}^- = \mathbf{x}_0^-$;
for $i = 1$ *to* m **do**
    $\mathbf{x}_b = (\mathbf{x}^+ + \mathbf{x}^-)/2$;
    Query f' for y_b, the label of $\mathbf{x}_b$;
    if y_b *is positive* **then**
        $\mathbf{x}^+ \leftarrow \mathbf{x}_b$
    else
        $\mathbf{x}^- \leftarrow \mathbf{x}_b$
    end
end

return $\{\mathbf{x}^+, \mathbf{x}^-\}$

---

2) Mid-perpendicular synthesis: There is an obvious limitation in binary search synthesis: if we always use binary search to generate training epochs, they may concentrate in one area and lack diversity. Therefore, we synthesize the next query along the mid-perpendicular direction after we find an opposite-pair close enough to the decision boundary. Specifically, we find the opposite-pair's orthogonal vector by Gram-Schmidt

process [32], set the magnitude of the orthogonal vector to q, then move it to a more precise midpoint. The details are shown in Algorithm 2.

Algorithm 2: Mid-perpendicular synthesis. $\mathbf{x}_s$ = $MidPerp(\{\mathbf{x}_b^+, \mathbf{x}_b^-\}, f', k, q)$

Input: $\{\mathbf{x}_b^+, \mathbf{x}_b^-\}$, an opposite-pair of EEG epochs; f', current substitute model; m, maximum number of binary search iterations; q, magnitude of the orthogonal vector.
Output: $\mathbf{x}_s$, an synthesized EEG epoch.

$\mathbf{x}_1 = \mathbf{x}_b^+ - \mathbf{x}_b^-$;
Generate an EEG epoch $\mathbf{x}_2$ randomly;
Find the orthogonal direction by Gram-Schmidt process:
$\mathbf{x}_2 = q \cdot (\mathbf{x}_2 - \langle \mathbf{x}_1, \mathbf{x}_2 \rangle / \langle \mathbf{x}_1, \mathbf{x}_2 \rangle \times \mathbf{x}_1)$;
$\{\mathbf{x}^+, \mathbf{x}^-\} = BinarySearch(\{\mathbf{x}_b^+, \mathbf{x}_b^-\}, f', m)$;
$\mathbf{x}_s = \mathbf{x}_2 + (\mathbf{x}^+ + \mathbf{x}^-)/2$.

return $\mathbf{x}_s$

With this query-syntheses-based active learning strategy, the entire substitute model training process for binary classification is shown in Algorithm 3.

Algorithm 3: Query-synthesis-based substitute model training strategy.

Input: f, the target model; S_0, a set of unlabeled EEG epochs; $N_{\max}$, maximum number of training epochs; $n_{\max}$, maximum number of synthesized EEG epochs in one iteration; m, maximum number of binary search iterations; q, magnitude of the orthogonal vector.
Output: f', a trained substitute model

Label S_0 by querying f to obtain an initial training set D;
Initialize f' and pre-train f' on D;
$\Delta D = \varnothing$;
for $N = 1$ *to* $N_{\max}$ **do**
    $\Delta S = \varnothing$;
    for $n = 1$ *to* $n_{\max}$ **do**
        Select an opposite pair $\mathbf{x}_0^+$ and $\mathbf{x}_0^-$ randomly from D;
        $\{\mathbf{x}_b^+, \mathbf{x}_b^-\} = BinarySearch(\{\mathbf{x}_0^+, \mathbf{x}_0^-\}, f', m)$;
        $x_s = MidPerp(\{\mathbf{x}_b^+, \mathbf{x}_b^-\}, f', m, q)$;
        $\Delta S \leftarrow \Delta S \bigcup \{\mathbf{x}_s\}$;
    end
    $\Delta D = \{(\mathbf{x}_i, f(\mathbf{x}_i))\}_{\mathbf{x}_i \in \Delta S}$;
    $D \leftarrow D \bigcup \Delta D$;
    Train f' on D;
end

return f'

We then extend query-synthesis-based augmentation method to multi-class classification, by the simple one-vs-one approach, which decomposes a multi-class task into multiple binary classification tasks. More specifically, if there are k_1 classes, then we can solve $k_2 = k_1(k_1 - 1)/2$ binary classifications instead, and use our active learning strategy to synthesize $n_{\max}/k_2$ EEG epochs (where $n_{\max}$ is the maximum number of synthesized EEG epochs in one iteration) for each binary task.

C. Adversarial Example Crafting for the Target Model

After training the substitute model, we can generate adversarial examples from it for the target model. Goodfellow *et al.* [15] proposed to construct adversarial perturbations in the following way:

$$\delta = \varepsilon \cdot \text{sign}(\nabla_{\mathbf{x}_i} J(\boldsymbol{\theta}, \mathbf{x}_i, y_i)), \tag{3}$$

where $\boldsymbol{\theta}$ are the parameters of the target model f, and J the loss function. The main idea is to find an optimal max-norm perturbation δ constrained by ε to maximize J. The requirements in (2) holds if $\varepsilon \leq \epsilon$ and l_∞-norm is used as the distance metric.

Let $\varepsilon = \epsilon$ so that we can perturb $\mathbf{x}_i$ at the maximum extent. Then, the adversarial example $\mathbf{x}_i^*$ can be re-expressed as:

$$\mathbf{x}_i^* = \mathbf{x}_i + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}_i} J(\boldsymbol{\theta}, \mathbf{x}_i, y_i)). \tag{4}$$

UFGSM [22] is an unsupervised extension of FGSM, which replaces the label y_i in (4) by $y_i' = f(\mathbf{x}_i)$. Then, $\mathbf{x}_i^*$ in UFGSM can be written as:

$$\mathbf{x}_i^* = \mathbf{x}_i + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}_i} J(\boldsymbol{\theta}, \mathbf{x}_i, y_i')). \tag{5}$$

UFGSM was used in this paper to construct the adversarial examples.

IV. EXPERIMENTS

This section presents the experimental results to demonstrate the effectiveness of the query-synthesis-based active learning strategy in transferability-based black-box attacks. We evaluate the vulnerability of three CNN classifiers in EEG-based BCIs.

A. Experimental Setup

The three BCI datasets, P300 evoked potentials (P300) [40], feedback error-related negativity (ERN)[1] [41], and motor imagery (MI)[2] [42], used in our recent study [22] were used again in this study. The data pre-processing steps were also identical. Three CNN classifiers, EEGNet [9], DeepCNN [10], and ShallowCNN [10] were used as the target models in our experiments, as in [22]. Adam optimizer [43], cross-entropy loss function, and early stopping were used in training. Moreover, we applied weights to different classes to address the class imbalance problem in P300 and ERN.

Raw classification accuracy (RCA) and balanced classification accuracy (BCA) were used to evaluate the attack performance, where RCA is the unweighted overall classification accuracy on the test set, and BCA is the average of the individual RCAs of different classes.

In order to simulate the black-box scenario, we partitioned the three datasets into two groups, as described in [22]: the larger group A of 7/14/7 subjects in P300/ERN/MI was used to simulate the unknown data from the target model, where 80% epochs were for training the target model and the remaining 20% for testing. The other smaller group B of 1/2/2 subjects was used to initialize set S_0 for training the substitute model. In this way, the 8/16/9 subjects in three datasets can be partitioned in 8/120/36 different ways. We performed both one-division and multi-division black-box attacks. In the one-division experiment, the attacks were repeated 5 times to reduce randomness. In the multi-division experiment, we repeated each division 5 times on P300 and only one time on ERN and MI; so, in total we had 40/120/36 evaluations on the three datasets, respectively.

B. Baseline

We first evaluated the baseline performance of the three CNN target models on the unperturbed EEG data, as shown in the first part of Table I. Because the MI dataset had 4 classes whereas P300 and ERN had only two, its RCAs and BCAs were much lower.

We then constructed a random perturbation δ':

$$\delta' = \epsilon \cdot \text{sign}\left(\mathcal{N}(0, 1)\right), \tag{6}$$

which has the same maximum amplitude ϵ as the adversarial perturbations, to verify the necessity of deliberately constructing the adversarial examples. The results are shown in the second part of Table I. It is obvious that the target models were robust to random noise, i.e., random noise cannot effectively perform adversarial attacks.

C. Attack Performance Comparison

We next compared our query-synthesis-based approach with the Jacobian-based method [22] in black-box attacks. $\epsilon = 0.1/0.1/0.05$ on P300/ERN/MI were used to construct the adversarial examples after the substitute models were trained.

In one-division experiments, we randomly downsampled the data for each class according to the labels that the target model predicted at the first time to balance the classes in the initial dataset. We set the *downsampled number* in each class as 200 in P300 and ERN, and 100 in MI, to ensure that the size of the initial substitute model training set S_0 was 400. The initial substitute model was then trained on S_0 using both the Jacobian-based method and ours. $\lambda = 0.5$ and $N = 1$ were used in the Jacobian-based method, where N corresponds to $N_{\max}$ in our method. $n_{\max} = 200$ and $m = 10$ were used in our method, so $N_{\max} = 2$ should be used to get the same number of queries as in the Jacobian-based method. $q = 1.0/0.8/0.8$ on P300/ERN/MI were used in the experiments. They were determined such that the generated

Experiment	Dataset	Target Model f	Baselines		Method	Substitute Model f'		
			Original	Noisy		EEGNet	DeepCNN	ShallowCNN
One-Division	P300	EEGNet	73.81/72.71	73.78/72.46	Ours	**54.90/54.63**	41.39/50.80	72.71/**71.79**
					Jacobian-based	62.36/61.64	54.58/55.55	**72.07**/72.02
		DeepCNN	77.77/74.79	78.08/74.78	Ours	69.77/70.52	**47.41/59.00**	**74.86/73.38**
					Jacobian-based	**66.40/66.82**	53.11/60.13	76.99/74.20
		ShallowCNN	72.56/72.87	72.64/72.87	Ours	**63.43/65.17**	**60.84/62.18**	**66.54/67.91**
					Jacobian-based	67.52/67.74	74.20/66.15	73.24/71.73
	ERN	EEGNet	76.26/74.19	75.47/73.34	Ours	**36.55/38.23**	**36.55/41.10**	**73.56/71.76**
					Jacobian-based	52.07/51.65	47.06/50.94	75.82/72.37
		DeepCNN	77.21/76.85	75.05/76.38	Ours	**46.22/46.57**	**52.07/61.59**	**75.14/73.99**
					Jacobian-based	51.65/52.79	55.36/64.33	75.67/75.97
		ShallowCNN	71.85/71.17	71.51/70.95	Ours	**70.69/71.44**	**69.75/70.66**	**44.82/49.81**
					Jacobian-based	72.16/71.50	70.73/71.22	52.10/53.57
	MI	EEGNet	54.93/54.62	52.93/52.78	Ours	**31.77/31.70**	35.96/36.07	**40.27/40.28**
					Jacobian-based	37.07/36.87	**35.26/35.24**	41.01/41.02
		DeepCNN	49.38/49.18	49.73/49.56	Ours	**41.91/41.78**	**39.24/39.08**	**38.14/38.05**
					Jacobian-based	45.36/45.22	40.56/40.48	41.09/41.06
		ShallowCNN	60.96/60.81	61.04/60.93	Ours	**51.93/52.06**	51.93/52.01	**43.60/43.68**
					Jacobian-based	55.05/55.18	**48.28/48.44**	48.97/49.00
Multi-Division	P300	EEGNet	73.59/71.95	73.52/71.83	Ours	**41.40/42.44**	**32.66/39.75**	**64.95/64.91**
					Jacobian-based	47.02/48.49	39.79/41.99	65.16/64.16
		DeepCNN	75.99/74.10	76.20/73.94	Ours	**53.29/55.10**	**36.57/44.55**	**68.78/66.88**
					Jacobian-based	59.18/60.01	44.54/49.66	70.05/66.98
		ShallowCNN	72.23/71.90	72.24/71.85	Ours	**59.66/60.17**	51.76/**55.15**	53.80/**49.60**
					Jacobian-based	59.75/61.73	**51.40**/57.91	**52.38**/52.54
	ERN	EEGNet	73.89/72.94	73.23/72.72	Ours	**45.71/47.29**	**46.78/48.94**	71.27/71.14
					Jacobian-based	54.42/54.47	52.54/54.53	**71.22/70.97**
		DeepCNN	74.24/72.69	73.86/72.38	Ours	**56.78/55.07**	**53.85/53.33**	**72.44/70.89**
					Jacobian-based	59.64/58.32	57.77/57.54	72.73/71.43
		ShallowCNN	71.86/71.45	71.77/71.21	Ours	**70.15**/70.46	**68.49/69.28**	**58.28/59.42**
					Jacobian-based	70.44/**70.30**	70.31/70.18	63.51/63.80
	MI	EEGNet	60.85/60.71	59.48/59.39	Ours	**35.10/35.27**	**46.47/46.45**	**43.60/43.67**
					Jacobian-based	39.13/39.13	46.58/46.61	53.19/53.17
		DeepCNN	55.83/55.57	55.65/55.39	Ours	**47.06/46.90**	**45.32/45.31**	**42.54/42.46**
					Jacobian-based	48.68/48.59	48.45/48.38	48.72/48.57
		ShallowCNN	64.71/64.62	64.18/64.09	Ours	**57.26/57.32**	**57.94/57.98**	**46.74/46.84**
					Jacobian-based	59.44/59.49	59.04/59.07	54.56/54.56

EEG epochs had approximately the same magnitudes as those in the initial training set S_0.

In multi-division experiments, we did not limit *downsampled number* to avoid wasting data, but always kept the same number of queries in the Jacobian-based method and ours. so $n_{max} = downsampling\ number$ on P300 and ERN, $n_{max} = 2 * downsampling\ number$ on MI in our method, the other hyper-parameters were the same as those in one-division experiments. The attack results are shown in Table I. We can observe that:

1) Generally, the RCAs and BCAs of the target models after attacks were lower than the corresponding baselines, indicating the effectiveness of the transferability-based black-box attack methods. However, shallowCNN did not perform well in attacking EEGNet and deepCNN on P300 and ERN.

2) The RCAs and BCAs after query-synthesis-based black-box attacks were generally lower than the corresponding accuracy after Jacobian-based black-box attacks. For example, in one-division experiments in which only 400 queries were used for both methods, when the target model and the substitute model were EEGNet and DeepCNN respectively on P300, our method achieved 13% improvement over the Jacobian-based method.

3) Our proposed method was often much better than the Jacobian-based method when the substitute model and the target model had the same structure, which may due to the strong intra-transferability of the adversarial examples.

We compared the attack performance with different number of queries when the substitute model and the target model had the same architecture, i.e., both were EEGNet, DeepCNN or
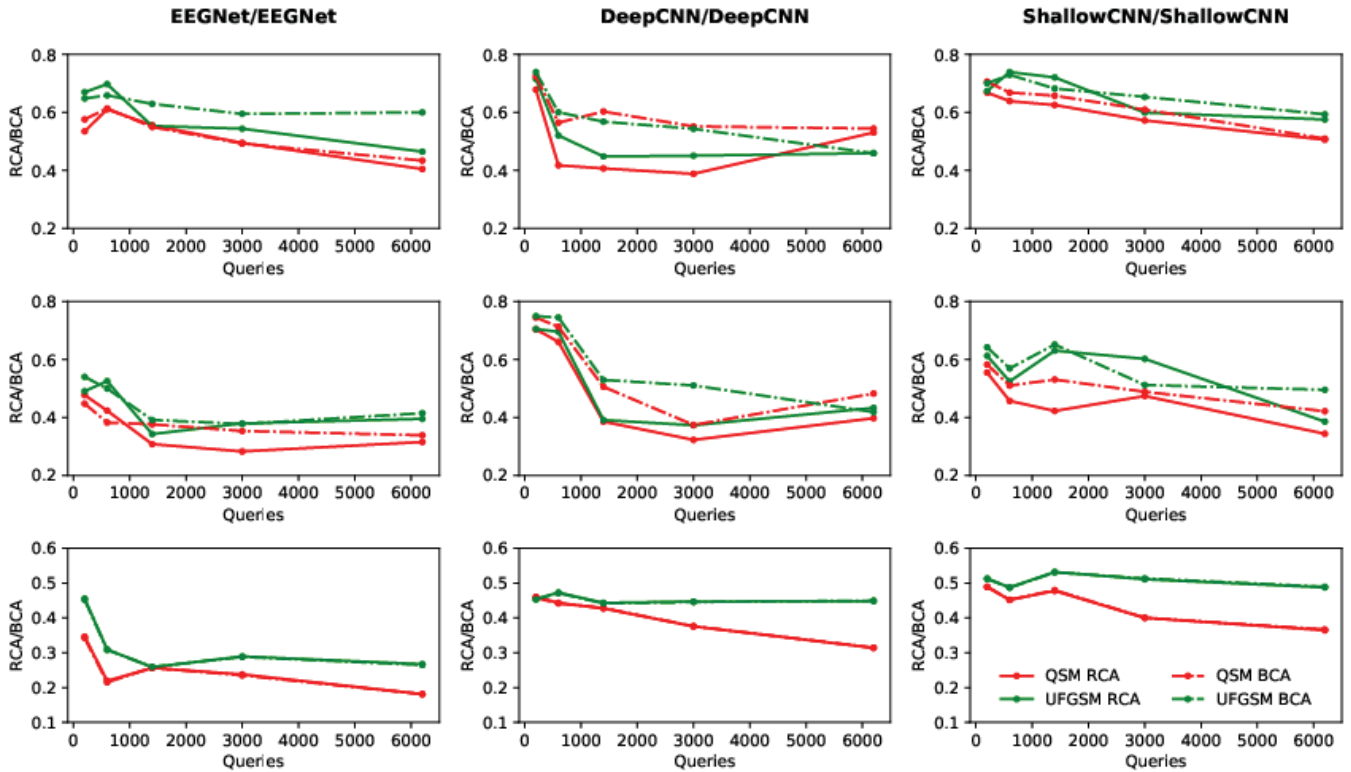
Fig. 1. RCAs and BCAs of the two black-box attack methods with different number of queries on P300 (top row), ERN (middle row) and MI (bottom row).

ShllowCNN. 200 EEG epochs were used for initializing the substitute model, and the number of queries varied from 200 to 6,200. The corresponding RCAs/BCAs are shown in Fig. 1. Generally, both RCAs and BCAs decreased as the number of queries increased, which is intuitive. With the same number of queries, our method achieved lower RCAs and BCAs than the Jacobian-based method, i.e., it gave higher attack success rates. In other words, it can achieve a desired attack success rate with much fewer queries. For example, for EEGNet on P300, our method only needed 3,000 queries to achieve a comparable attack performance with the Jacobian-based method with 6,200 queries.

D. Characteristics of the Adversarial Examples

In this section, we analyzed the characteristics of the generated adversarial examples in both time domain and spectral domain. Consider one-division query-synthesis-based black-box attacks on ERN after 400 queries, when both the target model and the substitute model are DeepCNN. An example of the original EEG epoch (we only show the first 10 channels) and its corresponding adversarial epoch is shown in Fig. 2. They were almost completely overlapping in the time domain, which means the adversarial example is very difficult to be detected by human or a computer. We had similar observations on other datasets and from other classifiers.

Next, spectrogram analysis was used to further explore the characteristics of the adversarial examples. Fig. 3 shows
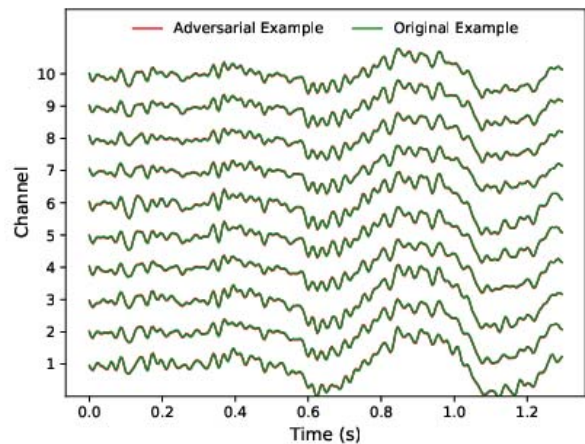

Fig. 2. Example of an original EEG epoch and its adversarial epoch, generated by query-synthesis-based black-box attack on the ERN dataset. The first 10 channels are shown. $\epsilon = 0.1$.

the mean spectrogram of all original EEG epochs, of all adversarial examples whose predicted labels did not match the ground-truth labels of the original EEG epochs (successful attacks), and of the corresponding successful perturbations $(\mathbf{x}^* - \mathbf{x})$, by using wavelet decomposition. The adversarial examples were crafted by query-synthesis-based black-box attacks on MI. Both the target and the substitute models
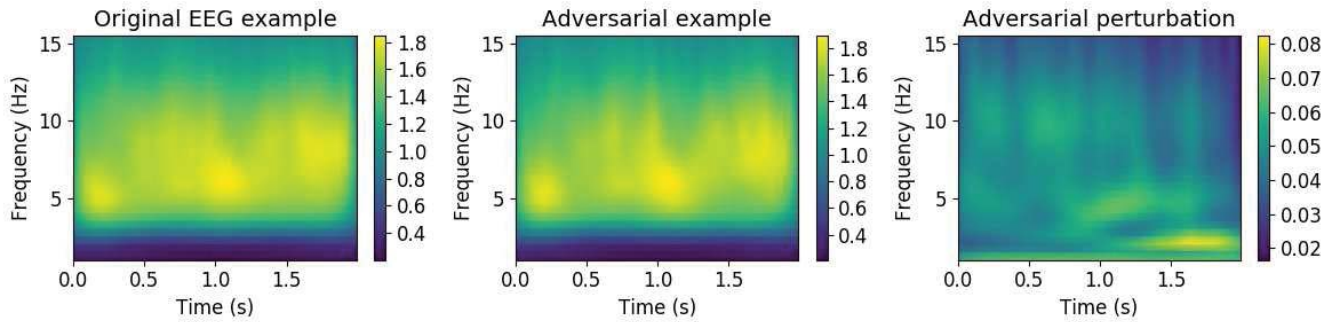
366

Fig. 3. Mean spectrogram of the original EEG epochs, of all successful adversarial examples, and of the corresponding perturbations, in one-division query-synthesis-based black-box attacks on the MI dataset. Channel F_z was used.

were DeepCNN. There was no significant difference in the spectrograms of the original EEG epochs and the adversarial examples. The amplitudes of the mean spectrogram of the adversarial perturbations were much smaller than those of the original and adversarial examples, suggesting that our generated adversarial examples are difficult to be detected by spectrogram analysis. We had similar observations on other datasets and from other classifiers.

V. CONCLUSION

In this paper, we have proposed a query-synthesis-based active learning strategy for transferability-based black-box attacks. It improves the query efficiency by actively synthesizing EEG epochs scattering around the decision boundary of the target model, and thus the trained substitute model can better approximate the target model. We applied our method to attacking three state-of-the-art deep learning classifiers in EEG-based BCIs. Experiments demonstrated its effectiveness and efficiency. With the same number of queries, it can achieve better attack performance than the traditional Jacobian-based method; or, in other words, our approaches needs a smaller number of queries to achieve a desired attack success rate.

REFERENCES

[1] B. Graimann, B. Allison, and G. Pfurtscheller, *Brain-Computer Interfaces: A Gentle Introduction*. Berlin, Heidelberg: Springer, 2009, pp. 1–27.

[2] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review," *Sensors*, vol. 12, no. 2, pp. 1211–1279, 2012.

[3] S. Sutton, M. Braren, J. Zubin, and E. R. John, "Evoked-potential correlates of stimulus uncertainty," *Science*, vol. 150, no. 3700, pp. 1187–1188, 1965.

[4] L. Farwell and E. Donchin, "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalography and Clinical Neurophysiology*, vol. 70, no. 6, pp. 510–523, 1988.

[5] D. Wu, "Online and offline domain adaptation for reducing BCI calibration effort," *IEEE Trans. on Human-Machine Systems*, vol. 47, no. 4, pp. 550–563, 2017.

[6] D. Wu, V. J. Lawhern, W. D. Hairston, and B. J. Lance, "Switching EEG headsets made easy: reducing offline calibration effort using active weighted adaptation regularization," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 11, pp. 1125–1137, 2016.

[7] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain-computer communication," *Proceedings of the IEEE*, vol. 89, no. 7, pp. 1123–1134, Jul 2001.

[8] D. Zhu, J. Bieger, G. Garcia Molina, and R. M. Aarts, "A survey of stimulation methods used in SSVEP-based BCIs," *Computational Intelligence and Neuroscience*, p. 702357, 2010.

[9] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, p. 056013, 2018.

[10] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[11] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," in *Proc. Int'l Conf. on Learning Representations*, San Juan, Puerto Rico, May 2016.

[12] Y. R. Tabar and U. Halici, "A novel deep learning approach for classification of EEG motor imagery signals," *Journal of Neural Engineering*, vol. 14, no. 1, p. 016003, 2017.

[13] Z. Tayeb, J. Fedjaev, N. Ghaboosi, C. Richter, L. Everding, X. Qu, Y. Wu, G. Cheng, and J. Conradt, "Validating deep neural networks for online decoding of motor imagery movements from EEG signals," *Sensors*, vol. 19, no. 1, p. 210, Jan. 2019.

[14] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition," in *Proc. ACM SIGSAC Conf. on Computer and Communications Security*. Vienna, Austria: ACM, Oct. 2016, pp. 1528–1540.

[15] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int'l Conf. on Learning Representations*, San Diego, CA, May 2015.

[16] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. Int'l Conf. on Learning Representations*, Banff, Canada, Apr. 2014.

[17] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proc. Int'l Conf. on Learning Representations*, Toulon, France, Apr. 2017.

[18] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proc. 35th Int'l Conf. on Machine Learning*, Stockholm, Sweden, Jul. 2018, pp. 284–293.

[19] N. Carlini and D. A. Wagner, "Audio adversarial examples: targeted attacks on speech-to-text," *CoRR*, vol. abs/1801.01944, 2018. [Online]. Available: http://arxiv.org/abs/1801.01944

[20] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel, "Adversarial perturbations against deep neural networks for malware classification," *CoRR*, vol. abs/1606.04435, 2016. [Online]. Available: https://arxiv.org/abs/1606.04435

[21] J. H. Metzen, M. C. Kumar, T. Brox, and V. Fischer, "Universal adversarial perturbations against semantic image segmentation," in *Proc. Int'l Conf. on Computer Vision*. Venice, Italy: IEEE, Oct. 2017, pp. 2774–2783.

[22] X. Zhang and D. Wu, "On the vulnerability of CNN classifiers in EEG-based BCIs," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 5, pp. 814–825, 2019.

[23] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symposium on Security and Privacy*. San Jose, CA: IEEE, May 2017, pp. 39–57.

[24] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks." in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*. Las Vegas, NV: IEEE, Jun. 2016, pp. 2574–2582.

[25] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proc. ACM Asia Conf. on Computer and Communications Security*. Abu Dhabi, UAE: ACM, Apr. 2017, pp. 506–519.

[26] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: reliable attacks against black-box machine learning models," in *Proc. Int'l Conf. on Learning Representations*, Vancouver, Canada, May 2018.

[27] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *Proc. Int'l Conf. on Machine Learning*, Stockholm, Sweden, Jul. 2018, pp. 2142–2151.

[28] A. N. Bhagoji, W. He, B. Li, and D. Song, "Practical black-box attacks on deep neural networks using efficient query mechanisms," in *Proc. European Conf. on Computer Vision*. Munich, Germany: Springer, Sep. 2018, pp. 158–174.

[29] D. Angluin, "Queries and concept learning," *Machine learning*, vol. 2, no. 4, pp. 319–342, 1988.

[30] R. D. King, J. Rowland, S. G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L. N. Soldatova *et al.*, "The automation of science," *Science*, vol. 324, no. 5923, pp. 85–89, 2009.

[31] I. Alabdulmohsin, X. Gao, and X. Zhang, "Efficient active learning of halfspaces via query synthesis," in *Proc. 29th AAAI Conf. on Artificial Intelligence*. Austin, Texas: AAAI Press, Jan. 2015, pp. 2483–2489.

[32] L. Wang, X. Hu, B. Yuan, and J. Lu, "Active learning via query synthesis and nearest neighbour search," *Neurocomputing*, vol. 147, pp. 426–434, 2015.

[33] L. Atlas, D. Cohn, R. Ladner, M. El-Sharkawi, R. Marks II, M. Aggoune, and D. Park, "Training connectionist networks with queries and selective sampling," in *Proc. Advances in Neural Information Processing Systems*. Denver, Colorado: MIT Press, Nov. 1989, pp. 566–573.

[34] S. Dasgupta, D. J. Hsu, and C. Monteleoni, "A general agnostic active learning algorithm," in *Proc. Advances in Neural Information Processing Systems*, British Columbia, Canada, Dec. 2008, pp. 353–360.

[35] B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," in *Proc. Conf. on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, Oct. 2008, pp. 1070–1079.

[36] D. Wu, V. J. Lawhern, S. Gordon, B. J. Lance, and C.-T. Lin, "Offline EEG-based driver drowsiness estimation using enhanced batch-mode active learning (EBMAL) for regression," in *Proc. IEEE Int'l Conf. on Systems, Man, and Cybernetics (SMC)*. Budapest, Hungary: IEEE, Oct. 2016, pp. 730–736.

[37] D. Wu, C.-T. Lin, and J. Huang, "Active learning for regression using greedy sampling," *Information Sciences*, vol. 474, pp. 90–105, 2019.

[38] D. Wu and J. Huang, "Affect estimation in 3D space using multi-task active learning for regression," *IEEE Trans. on Affective Computing*, 2019.

[39] E. B. Baum and K. Lang, "Query learning can work poorly when a human oracle is used," in *Proc. IEEE Int'l Joint Conf. on Neural Networks*, vol. 8, Baltimore, Maryland, Jun. 1992, p. 8.

[40] U. Hoffmann, J.-M. Vesin, T. Ebrahimi, and K. Diserens, "An efficient P300-based brain-computer interface for disabled subjects," *Journal of Neuroscience Methods*, vol. 167, no. 1, pp. 115–125, 2008.

[41] P. Margaux, M. Emmanuel, D. Sbastien, B. Olivier, and M. Jrmie, "Objective and subjective evaluation of online error correction during P300-based spelling," *Advances in Human-Computer Interaction*, vol. 2012, no. 578295, p. 13, 2012.

[42] M. Tangermann, K.-R. Mller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. Miller, G. Mueller-Putz, G. Nolte, G. Pfurtscheller, H. Preissl, G. Schalk, A. Schlgl, C. Vidaurre, S. Waldert, and B. Blankertz, "Review of the BCI Competition IV," *Frontiers in Neuroscience*, vol. 6, p. 55, 2012.

[43] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Proc. Int'l Conf. on Learning Representations*, Banff, Canada, Apr. 2014.

368